SENSITIVITY TO THE SERVICE-TIME DISTRIBUTION IN THE NONSTATIONARY ERLANG LOSS MODEL

by

Jimmie L. Davis,[1] *William A. Massey*[2] *and Ward Whitt*[3]

AT&T Bell Laboratories

October 20, 1992

Revision: May 17, 1993

[1] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332
(This work was done at AT&T Bell Laboratories as part of a CRFP fellowship.)

[2] AT&T Bell Laboratories, Room 2C-120, Murray Hill, NJ 07974-0636

[3] AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636

Abstract

The stationary Erlang loss model is a classic example of an insensitive queueing system: The steady-state distribution of the number of busy servers depends on the service-time distribution only through its mean. However, when the arrival process is a nonstationary Poisson process, the insensitivity property is lost. We develop a simple effective numerical algorithm for the M_t/PH/s/0 model with two service phases and a nonhomogeneous Poisson arrival process, and apply it to show that the time-dependent blocking probability with nonstationary input can be strongly influenced by the service-time distribution beyond the mean. With sinusoidal arrival rates, the peak blocking probability typically *increases* as the service-time distribution gets *less* variable. The influence of the service-time distribution, including this seemingly anomalous behavior, can be understood and predicted from the modified-offered-load and stationary-peakedness approximations, which exploit exact results for related infinite-server models.

Key Words: nonstationary queues; time-dependent arrival rates; nonhomogeneous Markov chains; transient behavior; Erlang loss model; blocking probability; insensitivity; infinite-server queues; modified-offered-load approximation.

1. Introduction and Summary

The classic Erlang loss model, which we denote by M/GI/s/0, has s homogeneous servers in parallel, no-extra waiting room, and independent and identically distributed (i.i.d.) service times that are independent of a Poisson arrival process. Arrivals finding all servers busy are lost, i.e., blocked without affecting future arrivals. Let S be a generic service time and let $G(t) \equiv P(S \leq t)$, $t \geq 0$, be its (general) cumulative distribution function (cdf). We assume that S has mean 1 and that the Poisson arrival process has rate α. Erlang (1918) showed that the steady-state blocking probability in this model is

$$B(s,\alpha) = (\alpha^s/s!)/\sum_{k=0}^{s} (\alpha^k/k!) \; ; \tag{1}$$

see Brockmeyer, Halstrom and Jensen (1960) and §3.3 of Cooper (1982).

Erlang recognized that the blocking probability in (1) depends on the service-time distribution only through its mean; indeed his primary argument was for the case of a deterministic service-time distribution. However, a proper proof was first provided by Sevastyanov (1957). This result is now part of a substantial insensitivity theory; e.g., see Chapter 6 of Franken et al. (1981) and Chapter 12 of Whittle (1986).

While the Erlang loss model has had many successful applications, it fails to represent the nonstationarity of real arrival processes. It has long been recognized that a much more realistic model is the M_t/GI/s/0 model, in which the arrival process is a *nonhomogeneous* Poisson process with deterministic arrival rate function $\alpha \equiv \alpha(t)$, but this model is much more difficult to analyze. Palm (1943) focused on the M_t/GI/s/0 model and investigated various ways to approximate the time-dependent blocking probability. (Throughout this paper the time-dependent blocking probability is defined as the time-dependent probability that all servers are busy.) Among other things, Palm noted that if the arrival rate changes slowly, then the stationary model with the instantaneous rate (i.e., the pointwise stationary approximation; e.g., see Green and Kolesar (1991) and Whitt (1991))) is approximately correct.

A common engineering approach is to use the stationary Erlang loss model with a constant arrival rate obtained as an average over an appropriate time interval during which the system is most heavily loaded, e.g., a busy hour. With this busy hour approach, the assumed arrival rate in the model is usually greater than or equal to the real arrival rate the majority of the time, so that the computed blocking probability tends to be conservative. This approach has been remarkably successful, especially when designing systems with a fixed number of servers that must be able to satisfy demand at any time. This approach is less successful for systems in which the number of servers varies dynamically, e.g., as with operator staffing. In order to determine the number of servers needed as a function of time, we want to be able to approximately determine the time-dependent blocking probability.

Analyses of time-dependent blocking usually assume exponential service times. This is in part obviously due to the exponential case being easier to analyze, but this focus may also be due in part to the well known insensitivity property that holds for the stationary model. The exponential distribution has also been shown to fit the data; indeed Erlang (1918) did statistical analysis showing that telephone call holding times are well approximated by an exponential distribution. However, telecommunications has evolved since then. Now the exponential distribution is not always considered appropriate. For example, with two or more classes of customers, each with exponential service times having different means, the overall service times have a hyperexponential distribution (a mixture of exponential distributions), which can be substantially more variable than an exponential distribution.

Hence, it is natural to ask whether the service-time distribution beyond its mean matters; i.e., it is natural to ask if insensitivity also holds in the nonstationary model. To investigate this question we consider the M_t/PH/s/0 model with a service-time distribution consisting of two phases; see Chapter 2 of Neuts (1981) for more on phase type (PH) distributions. This model allows us to treat exponential (M) distributions and both less-variable distributions such as two-phase Erlang (E_2) and more-variable distributions such as two-phase hyperexponential (H_2). Keeping track of the number of customers in each phase at any time, we obtain a finite-state continuous-time Markov chain (CTMC) with time-dependent infinitesimal generator. Following Koopman (1972), it is customary to numerically solve the ordinary differential equations (ODEs) corresponding to the Chapman-Kolmogorov equations by applying higher-order Runge-Kutta methods; see Gerald and Wheatley (1990), Taaffe and Ong (1987), Ong and Taaffe (1989) and Green, Kolesar and Svoronos (1991). However, instead we employ a direct discretization method, which produces a discrete-time Markov chain (DTMC) approximating the given CTMC. Since the CTMC can be defined as the limit of such DTMCs, the DTMC might well be considered the original model, so we are not particularly worried about the quality of the approximation. However, the approximations for the CTMC are also good. Our DTMC method is surprisingly effective (for admittedly relatively small problems). Our DTMC approach is similar to an Euler method for solving an ODE; e.g., see §5.2 of Gerald and Wheatley (1990). In §2 we describe the algorithm and discuss its performance. There we establish an error bound for the DTMC algorithm applied to the M_t/M/s/0 model. Our experience leads us to conclude that calculating exact distributions in models with time-varying rates need not be extraordinarily difficult.

Our numerical solution of the M_t/PH/s/0 model shows that insensitivity does not hold. Indeed, this conclusion can easily be deduced from analytical results for the $M_t/G/\infty$ model; see Eick, Massey and Whitt (1993a). However, since the pointwise stationary approximations based on (1) *does* have the insensitivity property, we might expect that the blocking probability in the M_t/G/s/0 model does not depend very much on the service-time distribution beyond its mean.

However, we find that *the service-time distribution beyond the mean can have a significant impact on the time-dependent blocking probability in the nonstationary Erlang loss model.* This effect is not great if the arrival rate changes very slowly (so that the model is nearly stationary at each point in time), but it can be significant when the arrival rate varies significantly in the time scale of individual service times.

Once we see that the service-time distribution beyond the mean can affect the time-dependent blocking probability in the nonstationary model, we might anticipate that greater variability produces more blocking, but this is not the case. In §3 we show for the case of sinusoidal arrival rates that *the peak blocking probability increases when the service-time variability (with fixed mean) decreases.*

In §4 we investigate the transient behavior of a stationary model starting out empty. More consistent with intuition, *steady-state is reached more slowly when the service-time distribution is more variable.*

In §5 we show that approximations based on associated infinite-server models can help explain the phenomena in §3 and §4. In particular, the observed influence of the service-time distribution is predicted, both qualitatively and quantitatively, by both the *modified-offered-load approximation* in Jagerman (1975) and Massey and Whitt (1993b) and the *stationary-peakedness approximation* in Massey and Whitt (1993c). This helps justify the attention we have recently given to infinite-server models in Eick, Massey and Whitt (1993a,b) and Massey and Whitt (1993a).

2. A Simple Effective Algorithm

By focusing on the number of busy server's in each phase of service in our M_t/PH/s/0 model, we obtain a finite-state CTMC with a time-dependent infinitesimal generator matrix $A(t)$. The time-dependent probability vector $\pi(t) \equiv (\pi_j(t)) \equiv \sum_i P(Q_s(t) = j | Q_s(0) = i) \pi_i(0)$ then can be obtained as the solution to the system of ordinary differential equations (ODEs)

$$\pi'(t) = \pi(t)A(t) \tag{2}$$

or, equivalently,

$$\pi(t) = \pi(0) + \int_0^t \pi(s)A(s)\,ds \ , \ t \geq 0 \ ; \tag{3}$$

see Dollard and Friedman (1979) for relevant theory. In general, we assume that $A(t)$ has nonpositive diagonal elements, nonnegative offdiagonal elements and zero row sums.

We approximate the CTMC by a DTMC and the integral (3) by the corresponding product. For this purpose, let

$$K = \sup_{0 \leq s \leq t} \{ |\max_i A_{ii}(s)| \} \ , \tag{4}$$

where $[0,t]$ is understood to be the time interval of interest. For step size $h > 0$, we define a DTMC with time-dependent transition function $P \equiv \{P(k) : k \geq 0\}$ with elements

$$P_{ij}(k) = \begin{cases} h\,A_{ij}(kh) \ , & i \neq j \\ 1 - \sum\limits_{\substack{m \\ m \neq i}} P_{im}(k) \ , & i = j \ . \end{cases} \tag{5}$$

To make sure that $P(k)$ is a bonafide transition matrix for each k, we require that the step size h in (5) be less than $1/K$ for K in (4). In practice, we choose h sufficiently small (less than $1/K$) so that the time-dependent state probabilities do not change substantially upon refinement.

We approximate $\pi(kh)$ by $\tilde{\pi}(kh)$, which we calculate recursively by

$$\tilde{\pi}(jh) = \tilde{\pi}((j-1)h)P(j-1) \ , \ j \geq 1 \ , \tag{6}$$

where $\tilde{\pi}(0) = \pi(0)$. Since $P(j)$ is sparse, the recursion (6) can be performed efficiently, e.g., we work with the positive elements of $P(j-1)$ instead of the matrices themselves.

It is customary to compute time-dependent state probabilities for time-dependent Markovian models by applying Runge-Kutta methods; e.g., Green, Kolesar and Svoronos (1991) report using fifth- and sixth-order Runge-Kutta methods in the International Math-Science Library subroutine DVERK. However, we found the simple DTMC method described here to be effective for the models we consider. Moreover, since CTMCs are defined as the limit of the DTMCs we consider, our approach seems very natural. We might well consider the approximating discrete-time chains to be the real model. We can also directly interpret probabilistically whether the step size is small enough; i.e., we can directly evaluate whether or not the DTMC seems to be a reasonable model.

Of course, it is also significant that the procedure is sufficiently accurate from the point of view of the CTMC. This is demonstrated by Table 1, which display a few numerical results for the $M_t/M/s/0$ model as a function of the step size *h*. In any example, the accuracy can easily be checked by refining the step size in this way (although this provides no guarantee).

To provide additional insight into this simple approach, we give an upper bound on the error for $M_t/M/s/r$ models, which is of order *h*. For this purpose, we use standard norms on vectors π and matrices *A*, defined by

$$|\pi| = \sum_i |\pi_i| \quad \text{and} \quad |A| = \max_i \{ \sum_j |A_{ij}| \} , \tag{7}$$

so that $|\pi A| \leq |\pi| \cdot |A|$ and $|A_1 A_2| \leq |A_1| \cdot |A_2|$. Let $\lambda'(t)$ be the derivative of $\lambda(t)$, which we assume is well defined when we use it.

Theorem. For an $M_t/M/s/r$ model with $\mu(t) = 1$ and $t = nh$,

$$\begin{aligned} |\pi(t) - \tilde{\pi}(t)| &\leq \sum_{k=1}^{n} [2h^2(\lambda(kh) + s)^2 + 2h \sup_{0 \leq u < h} \{|\lambda(kh) - \lambda(kh+u)|\}] \\ &\leq \eta \equiv 2h^2 \sum_{k=1}^{n} [(\lambda(kh) + s)^2 + \sup_{0 \leq u < h} \{|\lambda'(kh+u)|\}] , \end{aligned}$$

with

$$\eta \approx \tilde{\eta} \equiv 2h \int_0^t [(\lambda(u) + s)^2 + |\lambda'(u)|]\, du \ .$$

We prove this theorem in §6. This bound is not nearly tight, but it gives an indication of the way the error depends on the parameters.

3. The M_t/PH/s/0 Model With Periodic Arrival Rates

To study the effect of the service-time distribution on the time-dependent blocking, we consider two situations: dynamic steady state associated with a periodic arrival rate, see Heyman and Whitt (1984) and Thorisson (1985), and a one-time transient effect. To represent the first situation, we consider a sinusoidal arrival rate function in this section; to represent the second situation, we consider the transient behavior of a stationary model starting out empty in the next section.

Example 1. Consider an arrival-rate function of the form

$$\alpha(t) = \bar{\alpha} + \beta \sin(\gamma t) \ , \ t \geq 0 \ , \tag{8}$$

and let the system start empty at time 0. The parameters $\bar{\alpha}$, $\beta/\bar{\alpha}$ and γ are the *average arrival rate, relative amplitude* and *frequency*, respectively, see Eick et al. (1993b) for further discussion. Let $\bar{\alpha} = 10$, $\beta = 5$ and $\gamma = 1$.

We consider five different service-time distributions, all of which are special two-phase PH distributions; an Erlang (E_2), an exponential (*M*) and three hyperexponential (H_2) distributions. All have mean 1. The squared coefficients of variation (SCV, variance divided by the square of the mean, here the variance) are $c_s^2 = 0.5$ for E_2, $c_s^2 = 1.0$ for *M* and $c_s^2 = 4.0$ for the three H_2 distributions. (These are fixed for E_2 and *M*.) An H_2 density has the form

$$f(t) = p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t} \ , \ t \geq 0 \ , \tag{9}$$

so that there are three parameters. We used as a third parameter (in addition to a mean of 1 and

$c_s^2 = 4.0$)

$$r = p\lambda_1^{-1}/(p\lambda_1^{-1} + (1-p)\lambda_2^{-1}) \tag{10}$$

where $\lambda_1 < \lambda_2$; i.e., the proportion of the contribution to the mean provided by the component with the smaller λ_i. Formulas for p, λ_1 and λ_2 in terms of the mean, c_s^2 and r are given on p. 169 of Whitt (1984c) (Note that r here is $1-r$ there.) Here we consider three values of r: $0,1$, 0.5 and 0.9.

The time-dependent blocking probability calculated by our algorithm in §2 is shown in Figure 1. Note that dynamic steady state is achieved by the second cycle for the E_2 and M service-time distributions, but it takes somewhat longer for the H_2 distributions. (We focus on this issue more in §4.)

Since the maximum arrival rate is 15, the system is never exceptionally heavily loaded. If the arrival rate were fixed at its peak rate, then the stationary blocking probability would be $B(20,15) = 0.046$. In contrast, here the peak blocking ranges from about 0.01 for one of the H_2 distributions to about 0.035 for the E_2 distribution. On the other hand, if the arrival rate were fixed at its average rate, then the blocking probability would be only $B(20,10) = 0.0019$. The peak-rate stationary blocking and average-rate stationary blocking differ by a factor of about 24. (This is to show that the time-dependence makes a big difference. When these two stationary blocking probabilities do not differ much, there obviously is little to gain from careful analysis of the time-dependent behavior.)

In Figure 1, we see that *the peak time-dependent blocking probabilities for the different service-time distributions can differ by as much as a factor of 3.5*, so that the service-time distribution certainly can matter. We also see that the peak blocking probabilities are ordered according to the SCVs of the distributions, with the least variable distributions having the highest blocking.

From Figure 1, we also see that the service-time distribution beyond the first two moments also matter, because the three H_2 distributions yield quite different blocking. The blocking for the H_2 distributions is ordered according to the third parameter r. Increasing r, which corresponds to increasing the third moment, increases the peak blocking. (This is consistent with the known fact that, as the third moment increases, the H_2 renewal arrival process approaches a Poisson process. The H_2 renewal arrival process approaches a batch Poisson process as the third moment decreases; see Whitt (1984c).)

From Figure 1, we see that the peak blocking lags behind the peak arrival rate. Moreover, the lag is roughly independent of the service-time distribution.

The conclusions described so far hold consistently in other cases, although the difference in the time-dependent blocking is not always so great. In Figure 1 the less variable distributions produce higher time-dependent blocking almost uniformly over time. However, in Figure 1 the curves all coincide at their troughs, where there is nearly zero minimum blocking in each cycle. In other examples, the minimum blocking is not negligible. At the troughs, the distributions are also ordered, with the less variable distributions having lower minimum blocking probability. In other words, less variable distributions tend to produce greater fluctuations in their time-dependent blocking; they have higher peaks and lower troughs.

This phenomenon and the way the different cases approach steady state lead to a general physical explanation: *Systems with less variable service-time distributions tend to be more responsive to arrival-rate changes.* They will approach steady-state (dynamic or static) more quickly and achieve high or low blocking levels associated with new high or low arrival rates more quickly. Thus, less variable distributions should produce greater extremes with a periodic arrival process.

A simple way to see that more variable service times might reduce blocking is to consider a deterministic s-server loss system with arrivals at times $2n + j\varepsilon$, $1 \le j \le 2s$, for all positive integers n, where ε is very small so that $2s\varepsilon < 1$. If all the service times are exactly 1, then s customers are blocked in the interval $[2n, 2(n+1))$ for each n. However, if s of the $2s$ arrivals in $[2n, 2(n+1))$ have service time 2, while the remaining s have service time 0 (so that the average service time is still 1), then no customers at all are blocked.

The service-time distribution matters less as the frequency decreases. When we reduce the frequency γ from 1.0 to 0.2. the differences are much less than in Figure 1. The H_2 curves are close together with a peak blocking probability of about 0.03, while the exponential and Erlang curves are close together with a peak blocking probability of about 0.044. Now the big SCV $c_s^2 = 4.0$ of the H_2 distributions seems to be a key factor. Note that now the peak blocking for the E_2 and M distributions is closer to the stationary peak blocking probability $B(20,15) = 0.046$. To see that the influence of the service-time distribution still matters under higher loads, we consider $\bar{\alpha} = 15$, $\beta = 10$ and $\gamma = 1$. This produces results similar to Figure 1, but less dramatic. The peak blocking probabilities for the five service-time distributions here (in the same order) are about 0.28, 0.26, 0.24, 0.20 and 0.16.

Example 2. Our experimental evidence indicates that the influence of the service-time distribution beyond the mean tends to decrease as s increases. However, to show that the impact can be dramatic for very large s, we consider a simple (rather extreme) example. We let the arrival rate be periodic, but for simplicity non-sinusoidal. Let the arrival rate be $8s$ in the subintervals $[10k, 10k+1]$ for all k and let the arrival rate be 0 elsewhere. Let the mean service time be 1. It is easy to see that with deterministic service times the proportion of arriving customers blocked approaches 7/8 as $s \to \infty$, and the time-dependent blocking probability (probability all servers are busy) approaches 1 in the intervals $[10k+0.125, 10k+1]$ and 0 elsewhere as $s \to \infty$. In contrast, consider the two-point service-time distribution with $P(S = 9) = 1 - P(S = 0) = 1/9$, which also has mean 1. With this service-time distribution, the model is equivalent to one with a Poisson arrival process having rate $(8/9)s$ in the intervals $[10k, 10k+1]$ and 0 elsewhere, with deterministic service times of size 9. For this two-point service-time distribution, the proportion of customers blocked approaches 0 as $s \to \infty$. Moreover, the probability that all servers are busy approaches 0 uniformly in t as $s \to \infty$. While this example is not very realistic, it shows that the service-time effect need not go away when s becomes large.

To see what happens in a more realistic example with larger s, we considered the sinusoidal input in (8) again with $s = 100$. To obtain comparable experiments, we need to know how to choose the arrival-rate parameters for different s. Here heavy-traffic asymptotics as in Jagerman (1974) and Whitt (1984a) are helpful. In particular, the heavy-traffic asymptotics indicate that a relevant scaling is $\bar{\alpha}(s) = s + \sqrt{s} + o(\sqrt{s})$, $\beta(s) = b\sqrt{s} + o\sqrt{s}$ and $\gamma(s) = 0(1)$ as $s \to \infty$. Hence, to be similar to Example 1, we consider (2) with parameters $\bar{\alpha} = 90$, $\beta = 10$ and $\gamma = 1$. (This makes $\bar{\alpha} = s - \sqrt{s}$ and $\beta = -\sqrt{s}$.)

In this case, the service distribution beyond the mean still matters, but the differences are less dramatic than in Figure 1. The peak blocking probability ranges from about 0.05 to about 0.07. The slow convergence to dynamic steady-state for the H_2 distributions is especially evident.

4. Approach to Steady State Starting Empty

In this section we consider the stationary model with $\alpha(t) = \alpha$ for all $t \geq 0$, but not starting in steady state. In particular, we consider the approach to steady state starting out empty. For the case of exponential service times, the asymptotic behavior as $s \to \infty$ is investigated in Mitra and Weiss (1989).

Example 3. To relate to Example 1 consider the stationary Erlang loss model with $s = 20$ and $\alpha = \alpha(t) = 15$, $t \geq 0$, which has steady-state blocking probability $B(20, 15) = 0.046$. Figure 2 shows the time-dependent blocking probability for the five PH service-time distributions when the system starts out empty at time 0.

First observe that the celebrated insensitivity property is reflected by the convergence of all five curves to a common limit. Next observe that the service-time distribution has a significant impact. The service-time distributions are ordered as in §3; lower variability means more responsiveness; i.e., the approach to steady state is fastest with deterministic service times.

5. Insight Through Infinite-Server Approximations

We have seen that the service-time distribution can play an important role. Now it is natural to look for simple ways to explain and predict the effect. We suggest that infinite-server approximations can be very useful for this purpose. In this section we briefly discuss two infinite-server approximations and the insight that they can provide.

However, the most common approximation is probably the *pointwise-stationary approximation* (PSA). With PSA we approximate the time-dependent blocking probability at time t in the $M_t/G/s/0$ model by the stationary blocking formula (1) applied to the instantaneous offered load $\alpha(t)$, which is the arrival rate since $ES = 1$; i.e., the approximation at time t is $B(s, \alpha(t))$. It is important to note that the *PSA does not capture the service-time effect.* Since the stationary model has the insensitivity property, PSA produces a time-dependent blocking formula that is independent of the service-time distribution beyond its mean. From §3 and §4, we see that this is a shortcoming of PSA.

5.1 The Modified-Offered-Load Approximation

An alternative to PSA is the *modified-offered-load* (MOL) *approximation;* see Jagerman (1975) and Massey and Whitt (1993b). The idea is to approximate the time-dependent blocking probability by $B(s, m(t))$ instead of $B(s, \alpha(t))$, where $m(t)$ is the time-dependent mean number of busy servers in the associated $M_t/G/\infty$ infinite-server model with the same arrival-rate function and the same service-time distribution. Just like PSA, MOL is exact for the stationary model.

The number of busy servers at time t in the $M_t/G/\infty$ model has a Poisson distribution with mean

$$m(t) = E[\int_{t-S}^{t} \lambda(u)\,du] = E[\lambda(t-S_e)]E[S] , \tag{11}$$

where S_e is a random variable with the service-time stationary-excess distribution, i.e.,

$$P(S_e \le t) = \frac{1}{E[S]}\int_0^t P(S > u)\,du , \ t \ge 0 ; \tag{12}$$

see §1 of Eick et al. (1993a).

An important feature of the MOL approximation is that it *does* depend on the service-time distribution through the formula for the time-dependent mean $m(t)$ in (11). Moreover, the MOL approximation is reasonably accurate. As shown in Eick et al. (1993a,b), formula (11) is sufficiently tractable that we can learn a great deal from it about the way $m(t)$ depends on the arrival-rate function and, of particular interest here, on the service-time distribution. For example, explicit formulas for $m(t)$ for the models of §3 and §4 are given, respectively, in (7) of Eick et al. (1993b) and (20) of Eick et al. (1993a). For the case of service times with mean 1 and the sinusoidal input in (8),

$$m(t) = \bar{\alpha} + \beta \sin(\gamma t)E[\cos(\gamma S_e)] - \cos(\gamma t)E[\sin(\gamma S_e)] . \tag{13}$$

Formula (13) can be expressed conveniently in closed form in special cases, e.g., when S is exponential,

$$m(t) = \bar{\gamma} + \frac{\beta}{1+\gamma^2}[\sin\gamma t - \gamma\cos\gamma t] ; \tag{14}$$

see (15) of Eick et al. (1993b). Hyperexponential service is also treated explicitly in Eick et al. (1993b). For phase-type service, the $M_t/PH/\infty$ model is equivalent to a network of infinite-server queues, each with exponential service times, which is easily analyzed; see §8 of Massey and

Whitt (1993a). In particular, the vector of means (the mean number of busy servers in each service phase) is obtained as the solution of a k-dimensional ODE, where k is the number of service phases. Note that the computation is much less than for the exact solution in §2; there the ODE is $\binom{s}{k}$-dimensional, where s is the number of servers and k is again the number of service phases.

Similarly, for the transient behavior of the stationary model in §4,

$$m(t) = \alpha P(S_e \le t) \tag{15}$$

for S_e in (12). The first moment $ES_e = (c_s^2 + 1)/2$ is one useful representation of the relaxation time (time to approach equilibrium). Moreover, as indicated there, $m(t)$ decreases as the service-time distribution becomes more variable in the convex stochastic order.

When we look at the MOL approximations for the examples considered, we see that the phenomena observed in §2 and §3 are predicted, both quantitatively and qualitatively. This is illustrated by Figure 3, which displays the MOL approximations for Example 1. Figure 3 is not the same as Figure 1, but it is similar.

5.2 The Stationary-Peakedness Approximation

As discussed in Massey and Whitt (1993c) we can approximate the average blocking over an interval by approximation techniques for G/G/s/0 models, which have *stationary* non-Poisson arrival processes. Indeed, for the periodic arrival-rate functions considered in §2, we can regard the model as a stationary model with a non-Poisson arrival process if we randomize over a cycle.

It is natural to use approximations based on *peakedness.* The peakedness is the ratio of the variance to the mean of the steady-state number of busy servers in the associated G/G/∞ model; see Eckberg (1983), Whitt (1984a) and references there. In general, the peakedness is quite complicated, but there is a relatively simple approximation that is asymptotically correct as the arrival rate increases, namely,

$$z = 1 + (c_A^2 - 1) \int_0^\infty P(S > u)^2 \, du \tag{16}$$

and c_A^2 is the asymptotic variance of the arrival process; see p. 692 of Whitt (1984a).

Given that $\bar{\alpha} < s$, this approximation implies that *the average blocking probability should be increasing in the asymptotic peakedness* z in (16). Hence, we can gain insight by observing how z depends on the service-time cdf. Consistent with §3, we see that z *increases* as the service-time distribution gets *less* variable, provided that $c_A^2 > 1$. It is not difficult to show that indeed $c_A^2 > 1$ for the stationary version of a nonhomogeneous Poisson process. This somewhat counterintuitive behavior of stationary models is discussed in Wolff (1977) and in Remarks 6 and 7 of Whitt (1984b).

6. Proof of the Theorem

We conclude the paper by proving the theorem in Section 2. For stochastic matrices P_k and Q_k, it is easy to see that

$$\prod_{k=1}^{n} P_k - \prod_{k=1}^{n} Q_k = \sum_{k=1}^{n} \left[\prod_{i=1}^{k-1} P_i (P_k - Q_k) \prod_{j=k+1}^{n} Q_k \right] , \tag{17}$$

so that

$$| \prod_{k=1}^{n} P_k - \prod_{k=1}^{n} Q_k | \le \sum_{k=1}^{n} | P_k - Q_k | \ .$$

Similarly, let $E_A(t)$ and $E_B(t)$ are the time-dependent transition matrices associated with time-dependent generators $A(t)$ and $B(t)$; $E_A(t) = e^{At}$ when $A(t) = A$. Then

$$E_A(t) - E_B(t) = \int_0^t P_A(t)(A(s) - B(s)) E_B(t) \ ds$$

and

$$|E_A(t) - E_B(t)| \le \int_0^t |A(s) - B(s)|\, ds \ .$$

In addition, for any fixed generator matrix A

$$\exp(A) - I - A = \int_0^1 (1 - s) A^2 \exp(sA)\, ds \tag{18}$$

by Taylor's theorem with integral remainder. Since $\exp(sA)$ is a stochastic matrix, $|\exp(sA)| = 1$ and (18) implies that

$$|\exp(tA) - I - tA| \le \frac{t^2 |A|^2}{2} \ . \tag{19}$$

If A and $\tilde{A}$ are the generators of M/M/s/r systems with arrival rates λ and $\tilde{\lambda}$, and service rates 1 (not time-varying), then

$$|A| \le 2(\lambda + s) \tag{20}$$

and

$$|A - \tilde{A}| \le 2|\lambda - \tilde{\lambda}| \ .$$

Combining (19) and (20), we obtain

$$|\exp(hA) - I - hA| \le 2[(\lambda + s)h]^2 \ .$$

Now let $P_k = P(kh)$ in (14); let $Q_k = \exp(hA(kh))$; and let $B(t)$ be the time-varying generator defined by letting $B(s) = A(kh)$ for $kh \le s < (k+1)h$. Then

$$E_B(nh) = \prod_{k=1}^{n} Q_k \ . \tag{21}$$

Finally, combining these results, we obtain for $t = nh$

$$
\begin{aligned}
|\pi(t) - \tilde{\pi}(t)| &\le |E_A(t) - \prod_{k=1}^{n} P_k| \\
&\le |E_A(t) - E_B(t)| + |E_B(t) - \prod_{k=1}^{n} P_k| \\
&\le \int_0^t |A(s) - B(s)| ds + \sum_{k=1}^{n} |P_k - Q_k| \\
&\le \sum_{k=1}^{n} 2h \sup_{0 \le u \le h} \{ |\lambda(kh) - \lambda(kh+u)| \} + \sum_{k=1}^{n} 2h^2 (\lambda(kh) + s)^2 \ . \quad \blacksquare
\end{aligned}
$$

Acknowledgment. We thank Stephen G. Eick and Li-Chung Wang for their assistance in performing simulation experiments with the nonstationary loss model, which we have analyzed numerically here.

References

Brockmeyer, E., H. L. Halstrom and A. Jensen, *The Life and Works of A. K. Erlang,* second ed., *Acta Polytechnica Scandinavica Publishing Co.,* Stockholm, 1960.

Cooper, R. B., *Introduction to Queueing Theory,* second ed., North Holland, New York, 1982.

Dollard, J. D. and C. N. Friedman, *Product Integration with Applications to Differential Equations,* Encyclopedia of Mathematics and Its Applications, 10, Addison-Wesley, Reading, MA, 1979.

Eckberg, A. E., ‘‘Generalized Peakedness of Teletraffic Processes,’’ in *Proc. Tenth Int. Teletraffic Congress,* Montreal, Canada, 1983, paper 4.4.6.3.

Eick, S. G., W. A. Massey and W. Whitt, ‘‘The Physics of the $M_t/G/\infty$ Queue,’’ *Oper. Res.* 41 (1993a), to appear.

Eick, S. G., W. A. Massey and W. Whitt, ‘‘$M_t/G/\infty$ Queues with Sinusoidal Arrival Rates,’’ *Management Sci.* 39 (1993b), 241-252.

Erlang, A. K., ‘‘Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges,’’ *The Post Office Electrical Engineers’ Journal* 10 (1918), 189-197. (Translated from 1917 article in Danish in *Elektroteknikeren*, vol. 13).

Franken, P., D. König, U. Arndt and V. Schmidt, *Queues and Point Processes,* Akademie-Verlag, Berlin, 1981.

Gerald, C. F. and P. O. Wheatley, *Applied Numerical Analysis,* fourth ed., Addison-Wesley, Reading, MA, 1990.

Green, L. and P. Kolesar, ‘‘The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals,’’ *Management Sci.* 37 (1991), 84-97.

Green, L., P. Kolesar and A. Svoronos, ‘‘Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems,’’ *Oper. Res.* 39 (1991), 502-511.

Heyman, D. P. and W. Whitt, ‘‘The Asymptotic Behavior of Queues with Time-Varying Arrival Rates,’’ *J. Appl. Prob.* 21 (1984), 143-156.

Jagerman, D. L., ‘‘Some Properties of the Erlang Loss Functions,’’ *Bell System Tech. J.* 53 (1974), 525-551.

Jagerman, D. L., ‘‘Nonstationary Blocking in Telephone Traffic,’’ *Bell System Tech. J.* 54 (1975), 625-661.

Koopman, B. O., ‘‘Air Terminal Queues Under Time-Dependent Conditions,’’ *Oper. Res.* 20 (1972), 1089-1114.

Massey, W. A. and W. Whitt, ‘‘Networks of Infinite-Server Queues with Nonstationary Poisson Input,’’ *Queueing Systems* 12 (1993a), to appear.

Massey, W. A. and W. Whitt, ‘‘Analysis of the Modified-Offered-Load Approximation for the Nonstationary Erlang Loss Model,’’ 1993b, submitted.

Massey, W. A. and W. Whitt, ‘‘Stationary-Process Approximations for the Nonstationary Erlang Loss Model,’’ 1993c, submitted.

Mitra, D. and A. Weiss, ‘‘The Transient Behavior in Erlang’s Model for Large Trunk Groups and Various Traffic Conditions,’’ pp. 1367-1374 in *Teletraffic Science for the New Cost-Effective Systems, Networks and Services,* ITC-12, M. Bonatti (ed.), North-Holland, Amsterdam, 1989.

Neuts, M. F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach,* The Johns Hopkins University Press, Baltimore, 1981.

Ong, K. L. and M. R. Taaffe, ‘‘Nonstationary Queues with Interrupted Poisson Arrivals and Unreliable/Repairable Servers,’’ *Queueing Systems* 4 (1989) 27-46.

Palm, C., Intensity Variations in Telephone Traffic, *Ericsson Technics* 44 (1943), 1-189. (English translation by North-Holland, Amsterdam, 1988).

Sevastyanov, B. A., ‘‘An Ergodic Theorem for Markov Processes and its Application to

Telephone Systems with Refusals,'' *Theor. Probability Appl.* 2 (1957), 104-112.

Taaffe, M. R. and K. L. Ong, ''Approximating Ph(t)/M(t)/S/C Queueing Systems,'' *Ann. Oper. Res.* 8 (1987) 103-116.

Thorisson, H., ''On Regenerative and Ergodic Properties of the *k*-Server Queue with Nonstationary Poisson Arrivals,'' *J. Appl. Prob.* 22 (1985), 893-902.

Whitt, W., ''Heavy-Traffic Approximations for Service Systems with Blocking,'' *AT&T Bell Lab. Tech. J.* 63 (1984a), 689-708.

Whitt, W., ''Minimizing Delays in the GI/G/1 Queue,'' *Oper. Res.* 32 (1984b), 41-51.

Whitt, W., ''On Approximations for Queues, III: Mixtures of Exponential Distributions,'' *AT&T Bell Lab. Tech. J.* 63 (1984c), 163-175.

Whitt, W., ''The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues Is Asymptotically Correct as the Rates Increase,'' *Management Sci.* 37 (1991), 307-314.

Whittle, P., *Systems in Stochastic Equilibrium*, Wiley, New York, 1986.

Wolff, R. W., ''The Effect of Service Time Regularity on System Performance,'' in *Computer Performance*, pp. 297-304, K. M. Chandy and M. Reiser (eds.), North Holland, Amsterdam, 1977.

FIGURE 1. Exact numerical solutions for the $M_t/PH/s/0$ model starting empty in Example 1 with $s = 20, \quad \alpha(t) = 10 + 5\, SIN(t)$ and five service-time distributions.

FIGURE 2. Exact numerical solutions for the $M_t/PH/s/0$ model starting empty in Example 3 with $s = 20, \quad \alpha(t) = 15, \quad t \geq 0$ and five service-time distributions.

FIGURE 3. The modified-offered-load approximation for the $M_t/PH/s/0$ model starting empty in Example 1 with $s = 20, \quad \alpha(t) = 10 + 5\, SIN(t)$ and five service-time distributions.

Time t	Step Size		
	$h = 10^{-5}$	$h = 10^{-4}$	$h = 10^{-3}$
0.1	0.078543	0.078546	0.078576
2	0.189019	0.189017	0.188994
4	0.030160	0.030155	0.030111
6	0.020007	0.020001	0.020030
8	0.193267	0.193269	0.193285
10	0.047267	0.047260	0.047198
12	0.012346	0.012347	0.012355
14	0.186800	0.186806	0.186865
16	0.070326	0.070318	0.070241
18	0.008969	0.008969	0.008971
20	0.167856	0.167866	0.167969
accuracy		0.000010	0.000113

Table 1. The blocking probabilities in the $M_t/M/s/0$ model with $s = 100$ and $\lambda(t) = 100 + 20 \sin t$ calculated from the DTMC algorithm with various step sizes. The accuracy measures the maximum difference from the most refined case.

Time t	Step Size		
	$h=10^{-5}$	$h=10^{-4}$	$h=10^{-3}$
2.0	0.025060	0.025052	0.02497
2.5	0.262811	0.262882	0.26359
3.0	0.410136	0.410156	0.41036
3.5	0.459455	0.459462	0.45953
4.0	0.487940	0.487944	0.48798
4.5	0.503775	0.503777	0.50380
5.0	0.509244	0.509244	0.50925
5.5	0.505120	0.505118	0.50510
6.0	0.491032	0.491028	0.49099
6.5	0.465433	0.465427	0.46537
7.0	0.425283	0.425274	0.42519
7.5	0.365536	0.365523	0.36540
8.0	0.279290	0.279273	0.27910
8.5	0.163778	0.163757	0.16355
9.0	0.047881	0.047865	0.04771
9.5	0.001720	0.001718	0.00170
10.0	0.000000	0.000000	0.00000
accuracy		0.000071	0.00078

Table A1. The blocking probabilities in the $M_t/M/s/0$ model with $s=50$ and $\lambda(t) = 40t - 4t^2$ on $[0,10]$ calculated from the DTMC algorithm with various step sizes. The accuracy measures the maximum difference from the most refined case.

Time t	Step Size		
	$h = 10^{-5}$	$h = 10^{-4}$	$h = 10^{-3}$
0.1	0.002618	0.002619	0.002625
2.0	0.001810	0.001810	0.001812
4.0	0.003233	0.003233	0.003232
4.1	0.003271	0.003270	0.003260
4.2	0.001858	0.001857	0.001845
4.3	0.000933	0.000932	0.000927
4.4	0.000884	0.000884	0.000883
4.5	0.001564	0.001564	0.001566
4.6	0.002890	0.002890	0.002890
4.7	0.003463	0.003462	0.003455
4.8	0.002274	0.002272	0.002260
4.9	0.001090	0.001089	0.001081
5.0	0.000824	0.000824	0.000822
accuracy		0.000002	0.000014

Table 1. The blocking probabilities in the $M_t/M/s/0$ model with $s = 20$ and $\lambda(t) = 10 + 5 \sin 10t$ calculated from the DTMC algorithm with various step sizes. The accuracy measures the maximum difference from the most refined case.